ECE 595: Learning and Control
Paper #5 Summary

David Kirby

Due Friday, April 2, 2021 at 9:00 AM

Deep neural networks are vulnerable to adversarial attacks (small input perturbations that can drastically change output). There is the option to mitigate these perturbations by incorporating adversarial attacks during training, but this comes at the cost of excessive computational overhead due to iterative training. Incorporating the adversarial examples during training creates a min-max problem, whereby the adversary attempts to maximize the loss while the learner tries to minimize it. To reduce iterative training, this paper proposes interpreting the min-max problem as a robust optimal control problem, which will allow for deep neural network algorithms to optimize. By viewing the neural network as a discrete-time dynamical system, the authors suggest that we can view the min-max problem as a finite-horizon robust optimal control problem and reduce it using the Pontryagin Maximum Principle. This would allow us to create a Hamiltonian system that can be optimized using additive methods rather than the iterative and costly alternative. The authors' use of inexact gradient oracles is something I would like to discuss in class.